A global approach for a dictionary of Lingála

From the localization of software to a lemmatization strategy

Bienvenu Sene Mongaba

senemongaba@yahoo.fr Mabiki asbl/Universiteit Gent

Afrilex 5-7july 2011

1. Introduction

In DR Congo French is the vehicular language of teaching, but there is an insufficient command of it in today's school population (Manduku 2004 ; Nyembwe 2004 ; Saint Moulin 2009 ; Yawidi 2009 ; Sene M 2011a). This makes it desirable to use at least one of the four national languages of DR Congo (Lingála, Cilubà, Kikongo and Kiswahili) together with French, as a teaching strategy for better appropriation of knowledge and know-how.

In our sociolinguistics surveys, we observed that national languages are acquiring stronger and stronger empowerment in the linguistic market of DR Congo. This means that teachers are obliged to use national languages when they are teaching in order to be understood.

However, teachers are limited in using those languages because of the lack of teaching aids such as schoolbooks and dictionaries. Dictionaries are particularly important in this context, since sociolinguistic realities show that bilingual dictionaries are needed to explain in the national languages the meaning of the scientific terms which are used in French in the classroom.

Lingála is the national language spoken in Kinshasa, the capital of the country. Even if this language is going through a straightforward empowering process, i.e. it is used now in areas which in the past were the preserve of French, such as the media, justice and schools, its progress is still hindered by the lack of terminological and lexicographic tools; hence the need of providing monolingual or plurilingual dictionaries for this language, according to what is a demand of society.

This is why we have undertaken this research on lexicography. It consists in producing lexicons as teaching aids and in making them available to primary users, i.e. teachers and students.

In this communication, firstly, we will observe the approach followed by authors of existing Lingála dictionaries (monolingual and bilingual). This review aims to improve, if necessary, the presentation of entries, so as to help readers find the words they are looking for. The variations in the current spoken language (the so-called Kinshasa Lingála) and in the so-

called standard language (Makanza Lingála) will be important in the choice we will make for recording occurrences in the dictionary. This concerns the orthography of words, as well as lemmatized forms and register.

Secondly, we will describe the process of localization of software we will use to process the corpus.

Finally, we will also very briefly describe the corpus we are processing.

2. Existing Dictionaries

We started off observing what the traditional presentation is in existing dictionaries and we proposed some adaptations to achieve the best presentation in terms of contents, orthography and register.

At present, we have three main bilingual dictionaries and one monolingual dictionary:

Bilingual dictionaries:

Dzokanga (French-Lingála two volumes, Lingála-French, one volume).

Everbroeck (French-Lingála / Lingála-French: one volume)

Kawata (French-Lingála / Lingála-French: one volume)

Monolingual dictionary:

Kawata

These traditional dictionaries use the alphabetical presentation. Some observations about the presentation:

- ***The occurrence of verbs***

As all verbs begin by ko- in their infinitive form, the author chooses the lemmatized form beginning by the root.

Kotámbola will appear as *-támbola*

So, if one wants to look up the word ***kotámbola***, he must go to the section beginning by the letter ***t*** and he will find ***-támbola***.

- ***The occurrence of inflectional verbs***

Dictionaries don't list inflectional forms of verbs, but they list inflectional markers -mi-, -lo-,... The –mi- inflectional form occurs in the detail of a basic verb.

Example :

- mítánga (to rely on oneself) is listed at -tánga (count)

- ***The occurrence of derivative verbs***

Derivative forms of verbs are listed in dictionaries separately from the basic verb in alphabetical order.

Example : -*sála* (to work, to do, to make), -*sálisa* (to make someone work, to have someone do something or to help), -*sálama* (to be made), *sálana* (to do to each other)

- ***The occurrence of deverbative nouns***

Deverbative nouns are also listed separately:

Libála (marriage) is listed under the letter L while the basic verb *kobála* (to marry) is listed under the letter B as -*bála*

- ***The occurrence of single and plural forms of nouns***

There are no occurrences of plural forms for the classes 1-2, 9-2; 11-2 ; 14-2. For the classes 3-4 (mo-mi), 5-6 (li-ma) and 7-8 (e-bi), in general, only uncountable plurals or things which usually appear in large quantities (such as “beans”) are also listed in the plural form by some authors (Dzokanga). For words beginning by "e", the plural form is not listed except for *elóko* (thing), for which the plural *bilóko* (things) is listed (Everbroeck, Dzokanga). There are more occurrences for the couple li-ma. *Madésu* (beans), *mandéfu* (beard), *makaya* (cigarette), *makayabo* (salted fish). The singular form *libéle* (breast) is listed under the letter L and the plural form *mabéle* (breasts) is listed under the letter M because it also means milk.

Míso (eyes) and *míno* (teeth) are listed because they are similar exceptions to the li-ma rules. *Líso* (eye) should be *ma-íso*, the contraction of "*a″* and "í" gives *míso*. The same goes for *líno* (tooth): *ma-íno* becomes *míno*.

3. Lexicographic decisions

1. ***Issues of orthography***

The latest attempt at standardization and harmonization of writing of Congolese languages was made in the first conference of Congolese linguists in 1974. As Lingála is not taught in

the classroom and there are no grammar books written in Lingála for Lingála speakers, ordinary Lingála speakers write as they like. All of the variations described above make it more difficult for the lexicographer to choose one form of orthography. As we see above, the Lingála corpus reveals that there are different manners of writing a word. Let us see the main points.

1) Words can be written with their tones or not, but we choose to mark tones in our dictionary, as existing dictionaries do.
2) As to the fluctuation between "o" and "u" we mentioned above *(Mulúba or molúba),* we have chosen to consider the form with "o" as the correct form, because it is the original one, but we have listed the alternative form with "u" in alphabetical order, with a back-reference to the "o" form.
3) *Mái*, and *máyi* are two forms employed by users to write "water". We have chosen to consider the form "*mái*" as the correct form, because of the 1974 standardization, but we have listed the alternative form "*máyi*" in alphabetical order, with a back-reference to the "*mái*" form. The same approach has been adopted for all cases where alternative forms exist, with or without diphthongs :
   koúta and *kowúta*. We choose *koúta*
   nyónso, *niónso*, *nióso*, *nyóso*. We choose *niónso*

2. Related words and the prefix

How to show that words are related and provide the person who is looking for the word with a global explanation at the same time? In our presentation we opted for inserting an explanatory list of inflected and derived forms under the main lemma, with a back-reference under the derived forms, which are listed elsewhere in alphabetical order.

For example:

Kobála, kobálisa, kobálela, kobálana, libála, mobáli, babáli, mibáli, mobálani, babálani, bibálabála.

All of these words are related and their root is -*bál*-. So we will put at -*bál*- the whole explanation about those words. And when someone is looking for the meaning of *mobálani* which is a less frequent word according to our corpus, he will logically look it up in the alphabetical order. He will find *Mobálani*. There, the dictionary will tell him to look up -*bál*-. He will then go to *bál* and find *mobálani* in the alphabetical order of all words related to -*bál*-.

We also had to decide whether it would not be easier to put directly the explanation of *mobálani* at its general alphabetical occurrence and then put it again at *bál*. However, this solution would have led to having a dictionary of a bigger size, though more user-friendly.

3. The concordance rule of Bantu languages

We know that in Makanza Lingála the parts of speech respect the rule of concordance as is the case in Bantu languages. On the other hand, this rule is not used in Kinshasa Lingála and in fact it is not even understood by the majority of Lingála speakers, who as a rule use Kinshasa Lingála. We have chosen not to use the open-mid vowels ɛ and ɔ. However, we have chosen to mention concordance markers in a purely informative function.

4. The localization of software

1. Software localized

Localization means "that the entire lexicographic process, from initial compilation all the way to final product, may henceforth be conducted in any language of one's choice" (De Schryver 2006). Our work involved the localization of software in Lingála, which allowed us to work in a Lingála environment. The software programs concerned were Unitex, which provides linguistic resources and tools allowing us to analyze text in natural language, and TshwaneLex, which is a lexicographic software program allowing us to record data for the purpose of producing dictionaries.

2. Unitex

Localizing Unitex implies taking an existing language directory, copying it and renaming it in the chosen language. Then adaptations are made for the chosen language in a new directory. So, we used an English directory to create a Lingála directory. Localization consisted in adding the marked tone vowels in the alphabet file and the codes for noun classes, derivational verbs and specific Lingála tenses in the electronic dictionaries.

2.1. Alphabet

Localization began with the insertion of vowels not present in the alphabet file of the software. We incorporated the high tone vowels : á, é, í, ó, ú.

2.2. Electronic dictionaries (DELA)

a) The DELA Syntaxes

"The electronic dictionary distributed with Unitex uses the DELA syntax (LADL electronic dictionaries). This syntax describes the simple and compound lexical entries of a language with their grammatical, semantic and inflectional information (Unitex 2008:35). This is the dictionary created by the localizer to allow the annotation of words in the corpus. It's a text file written according to a set syntax allowing the software to recognize the inflected form of a given lemma.

Example of entry in the Unitex electronic dictionary (DELA):

```
Batatá, tatá.N+conc:/p/this is an example
```

This indicates that :

`Batatá` : inflected form. (the form we will find in the natural text).

`Tatá` : a canonic form (after lemmatization of `batatá`, Unitex identifies that "`tatá`" is the lemma)

The two forms are separated by a comma.

`N+conc` : the grammatical information. `Batatá` is a concrete Noun.

`p` indicates the plural form

"The dictionaries provided with Unitex contain descriptions of simple and compound words. These descriptions indicate the grammatical category of each entry, optionally their inflectional codes, and various semantic information" (Unitex 2008:39).

We adapted the English list provided, by adding some grammatical and semantic codes specific to Lingála.

b) Codes present in the English DELA

Code	**English description**
A	Adjective
ADV	Adverb
CONJC	Coordinating Conjunction
CONJS	Subordinating Conjunction
DET	Determiner
INT	Interjection
N	Noun
PREP	Preposition
PRO	Pronoun

V	Verb

Table 2: Frequent grammatical codes

Code	**English Description**
z1	General language
z2	Specialized language
z3	Very specialized language
Abst	Abstract
An1	Animal
An1coll	Collective animal
Conc	concrete
ConColl	Collective concrete
Hum	Human
HumColl	Collective Human
t	Transitive verb
i	Intransitive verb

Table 3: Some semantic codes

c) Codes for gender and number

Code	Description
~~m~~	~~Masculine~~
~~f~~	~~Feminine~~
~~n~~	~~Neuter~~
s	singular
p	Plural
1,2,3	1st, 2nd, 3rd person

Table 4: number codes. Words in Lingála don't have gender.

d) Codes for Lingála noun classes combination

Code Lingála noun classes
.N+mo-ba
.N+ø-ba
.N+mo-mi
.N+e-bi
.N+li-ma
.N+lo-ma
.N+lo-ba
.N+ki-bi
.N+ki-ba
.N+bo-ba

Table 6. Localization of Lingála noun classes. Common inflectional codes we have introduced

a) Codes for derived verbs and specific Lingála tenses.

Code Lingála tenses	**Tense affixes**	**Tense**	**Example**
P	-í	Elekí kala té	asálí
H	-aka	Momeseno	asálaka
C	-zalí koXa / -zá koXa / -zóXa / ákoXa	Ezalí koleka	azalí kosála / azá kosála / azósála / ákosála
B	-ákí	Eleká lisapo	asálákí
D	-á	Eleká etíkálá	asálá
E	-áká	Eleká kala	asáláká
G	-zalákí ko-	Eleká momeseno	azalákí kosála
F	-ko-	Ekoya	akosála
S	Γ- (ná-, ó-, á-, tó-, bó-, bá-, é-)	Pósá	násála, ósála, ásála, tósála, bósála, básála, ésála
I	~~É~~-á	etíndá	sálá
W	Ko-	Linoko	kosála
J	BoXáká	Loléngé ezalí koleka	Bosáláká
K	KoXaka	Linoko momeseno	Kosálaka

Table 7 localization of Lingála verb tenses codes

Code Lingála derivation		**mode**	**example**
L	-is-	causative	kosálisa
M	-el-	applicative	kosálela
Q	-am-	passive	kosálama
R	-an-	reciprocale	kosálana
T	-miX-	reflexive	komisála
TL	-miXis-		komisálisa
TM	-miXel-		komisálela
TQ	-miXam-		komisálama
LM	-isel-		kosálisela
LQ	-isam-		kosálisama
LR	-isan-		kosálisana
MR	-elan-		kosálelana
QM	-amel-		kosálamela

RL	-anis-		kosálanisa

Table 8 localization of Lingála derived verb codes

At the moment, we only insert possible double derivation. Some of these combinations are not permitted in Lingála. At this stage, we don't yet analyze triple combinations.

3. TshwaneLex

TshwaneLex allows to completely translate its commands into the language of the user. So we started out translating commands in Lingála. Generally, we used equivalent translation, creation of neologisms by derivation, compounding, and sometimes metaphoric naming.

Example :

a) equivalent translation

Create new list	Kelá molongó ya sika
Save ...	Bátélá ...
Search	Luká
Text	Makomi

b) creation of neologism by derivation (deverbative)

Entry Entries	Ekóta Bikóta
Entity Entities	Ezalisela Bizalisela

c) compounding

Alternatives	Basókíóyotéóyo
Database	Sandúkwabipésámí
Edit	BongísáBimísá
Activation	Sáláésála:
Frequency	Mbalaeyáka
Username	Kómbwámosáleli

Basókíóyotéóyo : plural of "if not this, this"

d) French loan adaptation

File	Fishé
Copy	Kopié
Code	Kóde
Percentage	Pursantaj

Letter	Létre to distinguish from nkomá reserved for characters)
Statistic	Sitatisitíki
Underline	Sulinyé

e) Loans without adaptation

Plugin	Plugin
Copyright	Copyright

f) Metaphoric

Fields	Mikala
Toggle bilingual Linked View	Kámbíká na etáleli ekangánisá minoko míbalé

5. THE CORPUS

1. Description of the corpus

Our corpus was constituted by:

1. Fiction and nonfiction books :
   - Fwa-ku-Mputu (Bienvenu Sene Mongaba, Ed Mabiki)
   - Bokobandela (Bienvenu Sene Mongaba, Ed Mabiki)
   - Mwana akimi ndako (Kando Taty Mbalaka, Ed Mabiki)
   - Mosuni (Esperance Bulayumi)
   - Ntoma (Stephanie Boale)
   - Bombula (Lemba Musalampasi)
   - Nalotoki Ndóto (Lemba Musalampasi)
   - Bamama ya Congo na Belgique (Bienvenu Sene Mongaba, Ed Mabiki)
   - The Holy Bible

2. Internet pdf documents

3. Internet chat fora

Sololabien, mbokamosika, congo2000

2. The corpus cleaning

The corpus is compiled in Notepad++ software. This software is useful for us because of its capacity to contain a big text file (>100Mo)

The majority of these texts do not mark tones. The first step of cleaning is to remove open-mid and diacritic signs. We will put them back later in a systematic way. However, the first cleaned corpus will be without those variations.

3. Final format

After cleaning the corpus, we saved it in Unicode format with openoffice. This is the format that the software Unitex can process.

4. Processing with Unitex

With Unitex, we extracted tokens and word frequency, collocations and made grammatical annotations.

Frequency is important in creating dictionaries. When creating a general dictionary, preference was given to words of high frequency while for specialized lexicons it was useful to look for these words among less frequent words.

Collocation helped us to find examples to illustrate our entries in the dictionary and to extract a possible definition from the corpus.

The annotation of the text allowed an early classification and lemmatization of the words contained in the corpus.

5. Importing data to TshwaneLex

Words extracted from the corpus and sorted with Unitex were then imported in TshwaneLex to create a bilingual (French and Lingála) general lexicon for primary and secondary school students. The work is still in progress.

6. Conclusions

This research has allowed us to localize Unitex according to Lingála grammar and to completely translate TshwaneLex. Working in a target language environment (Lingála) helped us have a global view of what we wanted to propose as an entry and its presentation. This is a global approach (lemmatization word and localization of software) we followed to produce a dictionary of Lingála.

7. References

De Schryver G-M 2006, *Internationalisation, Localisation and Customisation Aspects of TshwaneLex in* Lexikos 16 (AFRILEX-reeks/series 16: 2006) p. 223.

De Schryver G-M 2008, *A new way to lemmatize adjectives in a user-friendly Zulu-English Dictionary* Lexikos 18 (AFRILEX-reeks/series 18: 2008) 63-91.

Joffe D & De Schryver G-M 2009, *The TshwaneLex Suite. User Guide version 4.0.2.* http://tshanedje.com.

Dzokanga, *Dictionnaire sémantique illustré Français-Lingála* éd. Biso-moko.

Kawata A.T., 2003. *Bagó-Dictionnaire Lingála / Falansé Français-Lingála*, éd. Etóngá ya nkóta ya Kóngo

Kawata A.T., 2004. *Bagó ya Lingála ,* éd. Karthala.

Meeuwis M. 2010, *A grammatical overview of Lingála*, Lincom.

Paumier S. 2008, *Unitex 2.0 User manual.* http://www-igm.univ-mlv.fr/~unitex

Sene Mongaba 2006, *100 verbes pour parler Lingála*, Ed Mabiki.

Van Everbroeck R., 1985. *Maloba ma lokóta (Dictionnaire) Lingála Lingála-Français Français-Lingála*, éd. l'épiphanie.